NUI Galway
OÉ Gaillimh

Autumn Examinations 2021-2022

Course Instance Code(s)	4BCT1, 4BDS1, 4BMS2, 4BS2, 1MECE1, 1MEME1, SPE
Exam(s)	Fourth BSc in Computer Science & Information Technology Fourth Bachelor of Arts with Data Science Fourth Bachelor of Science (Mathematical Science) Fourth Bachelor of Science (Hons.) ME in Electronic & Computer Engineering ME in Mechanical Engineering Structured PhD
Module Code(s)	CT4101
Module(s)	Machine Learning
Paper No.	1

External Examiner(s)	Dr. Ramona Trestian
Internal Examiner(s)	Prof. Michael Madden *Dr. Patrick Mannion

Instructions:	Answer any 3 questions out of 4 questions. All questions carry equal marks (25 marks each). The total (out of 75 marks) will be converted to a percentage after marking.
Duration	2 hours
No. of Pages	5
Discipline(s)	School of Computer Science
Course Co-ordinator(s)	Dr. Colm O'Riordan (BCT), Dr. Nick Tosh (BDS), Prof. Dane Flannery (BMS), Dr. Emma Holahan (BS), Prof. Martin Glavin (MECE), Dr. Noel Harrison (MEME)

Requirements:

Release in Exam Venue	Yes [X]	No [ ]
MCQ Answersheet	Yes [ ]	No [X]
Handout	None	
Statistical/ Log Tables	None	
Cambridge Tables	None	
Graph Paper	None	
Log Graph Paper	None	
Other Materials	None	
Graphic material in colour	Yes [ ] No [X]	

PTO

Question 1 (25 marks)

Part (a)

Explain the differences between the following machine learning tasks: **classification**, **regression, clustering**. For each task type, list a specific application, and an algorithm which may be used to learn a suitable model.

[6]

Part (b)

Explain what is meant by the term **hypothesis function** in the context of a machine learning model. Provide an example of a hypothesis function for a regression problem with one independent variable, and provide a diagram illustrating your chosen hypothesis function and some example training data.

[5]

Part (c)

Explain how the **McCulloch and Pitts artificial neuron** works. You should include a labelled diagram of the neuron and all equations necessary to calculate the output of the neuron.

[6]

Part (d)

Explain what is meant by the term **loss function** in the context of artificial neural networks. Include an example of a loss function as part of your answer.

[4]

Part (e)

Briefly explain why machine learning practitioners use separate training and test datasets.

[4]

Question 2 (25 marks)

Part (a)

ID	Target	Prediction
1	FALSE	TRUE
2	FALSE	TRUE
3	FALSE	FALSE
4	TRUE	TRUE
5	FALSE	TRUE
6	FALSE	TRUE
7	TRUE	FALSE
8	FALSE	TRUE

The table above presents the results of evaluating a classifier on a test set for a binary classification task.

(i) Present the results of the evaluation above in a confusion matrix [3]
(ii) Calculate the accuracy of the classifier [2]
(iii) Calculate the true positive rate of the classifier [2]
(iv) Calculate the false negative rate of the classifier [2]
(v) Calculate the recall of the classifier [2]

Part (b)

Explain how receiver operating characteristic (ROC) curves may be used to compare the performance of different classifiers for a binary classification task. Sketch an example of an ROC curve for a binary classification task as part of your answer, clearly labelling the axes and clearly indicating a point that represents ideal performance.

[6]

Part (c)

Binning is a commonly used data pre-processing technique that allows a continuous feature to be converted into a categorical feature. Explain the difference between **equal frequency binning** and **equal width binning**. Include diagrams to aid your explanation.

[4]

Part (d)

Explain the difference between **correlation** and **causation**. Include a simple original example (i.e., one not covered in the lecture notes) to aid your explanation.

[4]

PTO

Question 3 (25 marks)

Part (a)

Describe the key similarities and key differences between the ***k*-means** algorithm and the ***k*-nearest neighbours** algorithm. As part of your answer provide a brief description of both algorithms.

[10]

Part (b)

Describe with the aid of a diagram how the elbow method may be used to identify a suitable value for *k* when applying the *k*-means algorithm to a dataset.

[4]

Part (c)

Explain with the aid of a diagram how a **Voronoi tessellation** may be used to create a visualisation of the decision boundaries for a 1-NN classification model.

[4]

Part (d)

Explain the differences between **range normalisation** and **z-normalisation**. As part of your answer, provide a simple example dataset that requires normalisation. Using one attribute from your example dataset, demonstrate how range normalised values and z-normalised values may be calculated.

[7]

Question 4 (25 marks)

ID	CHEST PAIN	BLOOD PRESSURE	HEART DISEASE
1	FALSE	HIGH	FALSE
2	TRUE	LOW	TRUE
3	FALSE	LOW	FALSE
4	TRUE	HIGH	TRUE
5	FALSE	HIGH	FALSE

The training dataset above contains data about the incidence of heart disease in 5 different patients. ID is a unique identifier for each data point, CHEST PAIN and BLOOD PRESSURE are independent attributes, and HEART DISEASE is the target attribute.

Part (a)
If you were to apply decision tree learning to the dataset above, which of the attributes would be the best choice for the root node of a decision tree? Justify your answer by providing complete calculations for information gain for the attribute CHEST PAIN and the attribute BLOOD PRESSURE.

[12]

Part (b)
If you were to apply the k-nearest neighbours algorithm to the dataset above, what would be a suitable choice of distance metric? Provide a brief justification for your answer.

[2]

Part (c)
Describe in detail (including pseudocode) an algorithm that could be used to learn a decision tree for a classification task.

[7]

Part (e)
Briefly describe how you would detect overfitting in a decision tree classifier. As part of your answer briefly outline a technique that could be used to combat overfitting in a decision tree.

[4]

END OF EXAM PAPER